ANIMAL GENETICS Immunogenetics, Molecular Genetics and Functional Genomics

SHORT COMMUNICATION doi:10.1111/j.1365-2052.2009.02008.x

Discovery, validation and characterization of 1039 cattle single nucleotide polymorphisms

R. Donthu*, D. M. Larkin*, M. P. Heaton† and H. A. Lewin*,‡

*Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. †United States Department of Agriculture, Agricultural Research Service, US Meat Animal Research Center, Clay Center, NE 68933, USA. ‡Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA

Summary

We identified ~13 000 putative single nucleotide polymorphisms (SNPs) by comparison of repeat-masked BAC-end sequences from the cattle RPCI-42 BAC library with whole-genome shotgun contigs of cattle genome assembly Btau 1.0. Genotyping of a subset of these SNPs was performed on a panel containing 186 DNA samples from 18 cattle breeds including 43 trios. Of 1039 SNPs confirmed as polymorphic in the panel, 998 had minor allele frequency ≥0.25 among unrelated individuals of at least one breed. When Btau 4.0 became available, 974 of these validated SNPs were assigned *in silico* to known cattle chromosomes, while 41 SNPs were mapped to unassigned sequence scaffolds, yielding one SNP every ~3 Mbp on average. Twenty-four SNPs identified in Btau 1.0 were not mapped to Btau 4.0. Of the 1015 SNPs mapped to Btau 4.0, 959 SNPs had nucleotide bases identical in Btau 4.0 and Btau 1.0 contigs, whereas 56 bases were changed, resulting in the loss of the *in silico* SNP in Btau 4.0. Because these 1039 SNPs were all directly confirmed by genotyping on the multi-breed panel, it is likely that the original polymorphisms were correctly identified. The 1039 validated SNPs identified in this study represent a new and useful resource for genome-wide association studies and applications in animal breeding.

Keywords single nucleotide polymorphism discovery, cattle breeds.

Discovery of single nucleotide polymorphisms (SNPs) has facilitated the characterization of linkage disequilibrium and fine-mapping of quantitative trait loci (QTL) in cattle (Khatkar *et al.* 2007; Daetwyler *et al.* 2008). More than 2 million *in silico* SNPs were detected in the cattle genome using whole-genome shotgun (WGS) reads (Bovine Genome Sequencing and Analysis Consortium *et al.* 2009). Another publicly available SNP source for cattle was developed from the clustering and alignments of cattle expressed sequence tags (ESTs) to the consensus sequences (Hawken *et al.* 2004). In the SNP discovery process, Hawken *et al.* (2004) did not utilize EST sequence trace files. As a result, these SNPs tend to have a lower validation rate than SNPs detected using WGS reads (Gautier *et al.* 2007). Recently, Van Tassell *et al.* (2008) discovered ~60 000 SNPs using deep sequencing of reduced representation libraries. In three cattle populations, these SNPs had higher minor allele frequencies compared with SNPs identified from WGS reads. The aim of the present study was to identify SNPs from BAC-end sequences (BESs) and to validate them on 18 cattle breeds.

BAC-end sequences obtained from the RPCI-42 BAC library (Holstein) were downloaded from GenBank. Repeat-masked BESs were compared with the WGS contigs of cattle genome assembly (Btau 1.0) using the TimeLogic TERA-BLASTN program ($E < 10^{-50}$) (Fig. 1). All BLASTN alignments were parsed using custom Perl scripts and only the alignments (with a minimum length of 100 bp) containing <15 mismatches between BESs and WGS contigs were used for the identification of *in silico* SNPs. All the single base mismatches that had perfect alignments in the 10 flanking bases on either side were considered as putative SNPs. For all putative SNPs and the 10 flanking bases on either side of the SNP position, Phred quality scores (Q) (Ewing *et al.* 1998) were retrieved from a local database that was created for the BESs (Everts-van der Wind *et al.* 2005). A putative SNP was classified as a high-quality *in silico* SNP if the base at the SNP position in a BES and WGS contig had $Q \geq 30$ and $Q \geq 60$ respectively, and the 10 flanking bases on either side in the same BES and WGS contig had $Q \geq 20$ and $Q \geq 40$ respectively.

Address for correspondence

H. A. Lewin, Institute for Genomic Biology and Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois, 61801, USA.
E-mail: h-lewin@illinois.edu

Accepted for publication 29 October 2009

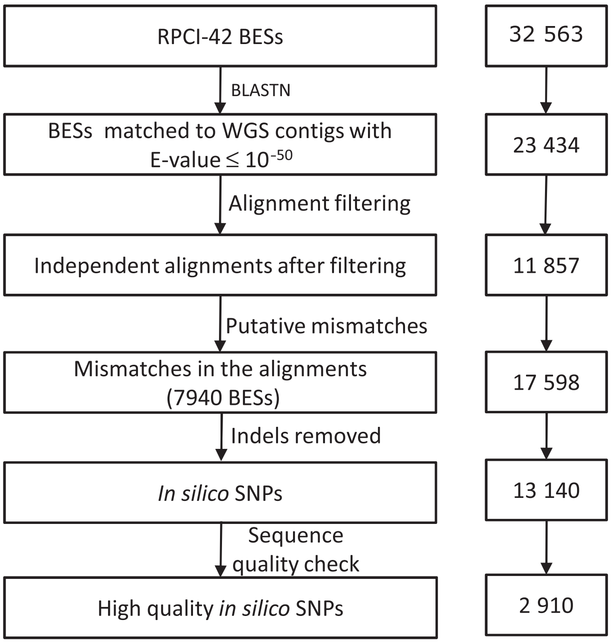
RPCI-42 BESs
32 563
BLASTN
BESs matched to WGS contigs with E-value ≤ 10-50
23 434
Alignment filtering
Independent alignments after filtering
11 857
Putative mismatches
Mismatches in the alignments (7940 BESs)
17 598
Indels removed
In silico SNPs
13 140
Sequence quality check
High quality in silico SNPs
2 910

Figure 1 Schema of *in silico* SNP discovery. The schema shows intermediate output at each step of the process.

Table 1 Distribution of SNPs within different cattle breeds.

Breed	No. of samples[1]	No. of SNPs polymorphic	No. of SNPs with MAF ≥0.25
Angus	8	832	471
Beefmaster	5	805	466
Brahman	6	554	253
Brangus	5	791	442
Charolais	6	849	537
Chianina	4	737	509
Gelbvieh	17 (12)	892	374
Hereford	8	879	578
Holstein	4	796	581
Limousin	8	873	525
Longhorn	4	739	533
Maine–Anjou	5	790	429
Red Angus	6	810	517
Simmental	24 (16)	911	468
Salers	5	761	415
Santa Gertrudis	4	750	512
Shorthorn	5	770	405
South Devon	62 (24)	952	502
All breeds	186	1039	540

[1] Number in parenthesis is the number of unrelated individuals in the breed used for estimating MAF. Among the 43 trios from the IRRF used for estimating MAF, there were six unique sires, 37 unique dams and 43 offspring, for a total of 86 samples. The 100 additional samples were 96 from the USDA-MARC beef cattle diversity panel and four controls of the South Devon breed.
MAF, minor allele frequency.

A total of 17 598 mismatches were identified in the alignments of 7940 BESs to WGS contigs. After removing indels, there were 13 140 *in silico* SNPs, of which 2910 passed the threshold for high quality (Fig. 1). We selected 814 high-quality *in silico* SNPs and 2258 from the remaining *in silico* SNPs for genotyping on the Illumina BeadArray platform (Oliphant *et al.* 2002) using a panel of 186 samples from 18 cattle breeds (Table 1). The panel includes 43 trios from the Illinois Reference and Resource Families (IRRF: Ma *et al.* 1996) and 96 samples from the USDA Meat Animal Research Center beef cattle diversity panel version 2.1 (Heaton *et al.* 2001).

A total of 763/814 (93.7%) high-quality *in silico* SNPs and 276/2258 (12.2%) unfiltered *in silico* SNPs were found to be polymorphic in at least one of the genotyped individuals. Confirmation of 93.7% high-quality *in silico* SNPs showed the effectiveness of our SNP discovery strategy and also revealed the importance of checking the quality of sequence bases in SNP discovery. Validation of 12.2% unfiltered *in silico* SNPs demonstrates that such SNPs are a potentially rich but unreliable resource. All 1039 verified SNPs were submitted to the NCBI dbSNP database (accession numbers can be found in Table S1) and a custom track made available on the UCSC Genome Browser (Kent *et al.* 2002) (http://www-app.igb.uiuc.edu/labs/lewin/donthu/CT/html/). The frequency of scorable genotypes (average call rate) was 95.5%, and correct inheritance in sample trios was 99.5%, thus demonstrating the high accuracy of the genotyping method.

Of 1039 confirmed SNPs, 998 had minor allele frequency (MAF) of ≥0.25 in at least one breed when only unrelated individuals were analysed (Table 1). However, among the 18 cattle breeds studied, only Simmental, Gelbvieh and South Devon had sample sizes sufficient for reliable estimates of MAF. In these breeds, South Devon had the maximum proportion of SNPs with MAF >0.40, while Gelbvieh had none (Fig. 2). Although sample size was relatively small for Brahman, it had the fewest SNPs, with MAF ≥0.25 of all breeds tested (Table 1). There were five SNPs (rs41256698, rs41256247, rs41256704, rs41256261 and rs41257796) with MAF ≥0.25 that are shared by all tested breeds.

When cattle genome assembly Btau 4.0 became available, we compared flanking bases of all validated SNPs against the

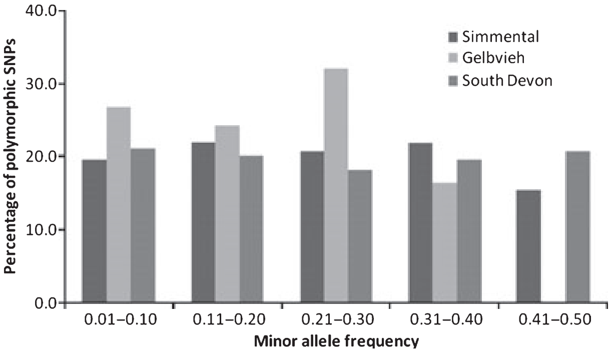

Figure 2 Distribution of minor allele frequencies within all breeds with sample sizes adequate for a robust estimation of minor allele frequency.

doi:10.1111/j.1365-2052.2009.02008.x

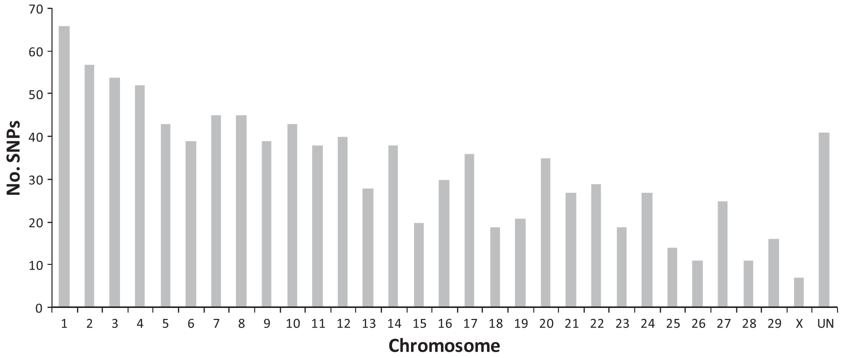

Figure 3 Distribution of SNPs on cattle chromosomes and unassigned contigs (UN).

chromosome sequences using TERA-BLASTN ($E < 10^{-50}$) to obtain the SNP position in the current assembly. We assigned the positions of 974 verified SNPs to known chromosomes, while an additional 41 validated SNPs were mapped to unassigned sequence scaffolds (Fig. 3; Table S1). Twenty-four SNPs identified in Btau 1.0 were not mapped to Btau 4.0. The bases at the SNP positions were identical for 959 of 1015 SNPs in both Btau 1.0 and Btau 4.0 contigs. However, the bases were different in both Btau 1.0 and Btau 4.0 contigs at 56 SNP positions, suggesting that there is no SNP at these positions in Btau 4.0. As these SNPs were validated on a multi-breed panel in the present study, it is likely that polymorphisms exist at these positions. However, because of the process used for contig assembly, it is possible to find SNPs in different sequence assemblies at homologous sites. This suggests that when using only the reference sequence for SNP search, many SNPs may be missed. Even if the reference sequence is derived from a single individual, all sequence reads should be included in the alignment and subsequent SNP searches. The set of *in silico* and confirmed SNPs we have identified thus forms a reliable resource for genetic analysis within and among cattle breeds.

Acknowledgments

This study was supported in part by funding from USDA-ARS contract No. 538 AG 58-54387-317 and USDA-CSREES Special Grant No. 538 AG 2006-34480-17150.

References

Bovine Genome Sequencing and Analysis Consortium, Elsik C.G, Tellam R.L. & Worley K.C. (2009) The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science* **324**, 522–8.

Daetwyler H.D., Schenkel F.S., Sargolzaei M. & Robinson J.A.B. (2008) A genome scan to detect quantitative trait loci for economically important traits in Holstein cattle using two methods and a dense single nucleotide polymorphism map. *Journal of Dairy Science* **91**, 3225–36.

Everts-van der Wind A., Larkin D.M., Green C.A., Elliott J.S., Olmstead C.A., Chiu R., Schein J.E., Marra M.A., Womack J.E. & Lewin H.A. (2005) A high-resolution whole-genome cattle–human comparative map reveals details of mammalian chromosome evolution. *Proceedings of National Academy of Sciences of the United States of America* **102**, 18526–31.

Ewing B., Hillier L.D., Wendl M.C. & Green P. (1998) Base-calling of automated sequencer traces using *Phred*. I. Accuracy assessment. *Genome Research* **8**, 175–85.

Gautier M., Faraut T., Moazami-Goudarzi K. *et al.* (2007) Genetic and haplotypic structure in 14 European and African cattle breeds. *Genetics* **177**, 1059–70.

Hawken R.J., Barris W.C., McWilliam S.M. & Dalrymple B.P. (2004) An interactive bovine *in silico* SNP database (IBISS). *Mammalian Genome* **15**, 819–27.

Heaton M.P., Chitko-McKown C.G., Grosse W.M., Keele J.W., Keen J.E. & Laegreid W.W. (2001) *Interleukin-8* haplotype structure from nucleotide sequence variation in commercial populations of U.S. beef cattle. *Mammalian Genome* **12**, 219–26.

Kent W.J., Sugnet C.W., Furey T.S., Roskin K.M., Pringle T.H., Zahler A.M. & Haussler D. (2002) The human genome browser at UCSC. *Genome Research* **12**, 996–1006.

Khatkar M.S., Zenger K.R., Hobbs M. *et al.* (2007) A primary assembly of a bovine haplotype block map based on a 15,036-single-nucleotide polymorphism panel genotyped in Holstein-Friesian cattle. *Genetics* **176**, 763–72.

Ma R.Z., Beever J.E., Da Y. *et al.* (1996) A male linkage map of the cattle (*Bos taurus*) genome. *Journal of Heredity* **87**, 261–71.

Oliphant A., Barker D.L., Stuelpnagel J.R. & Chee M.S. (2002) BeadArray technology: Enabling an accurate, cost-effective approach to high-throughput genotyping. *BioTechniques* 32, 56–61.

Van Tassell C.P., Smith T.P.L., Matukumalli L.K., Taylor J.F., Schnabel R.D., Lawley C.T., Haudenschild C.D., Moore S.S., Warren W.C. & Sonstegard T.S. (2008) SNP discovery and allele frequency estimation by deep sequencing of reduced representation libraries. *Nature Methods* **5**, 247–52.

Supporting information

Additional supporting information may be found in the online version of this article.

Table S1 Minor allele frequencies of 1039 validated SNPs in 18 cattle breeds.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

doi:10.1111/j.1365-2052.2009.02008.x